AD NUMBER

AD460819

NEW LIMITATION CHANGE

TO

Approved for public release, distribution unlimited

FROM

Distribution authorized to U.S. Gov't. agencies and their contractors; Administrative/Operational Use; 23 NOV 1964. Other requests shall be referred to Chief of Research and Development [Army], Washington, DC.

AUTHORITY

USAMC ltr, 23 May 1967

AD

U.S. ARMY

CHEMICAL INFORMATION & DATA SYSTEM

A COLLECTION OF ALGORITHMS FOR SEARCHING CHEMICAL COMPOUND STRUCTURE ANALOGS

by
WILLIAM J. WILSON & JOHN B. BURGER

CIDS Report No. 3
NOVEMBER 1964
PREPARED AT REDSTONE SCIENTIFIC INFORMATION CENTER

DIRECTOR OF ARMY TECHNICAL INFORMATION
OFFICE OF CHIEF RESEARCH & DEVELOPMENT
DEPARTMENT OF THE ARMY

DDC AVAILABILITY NOTICE

All distribution of this report is controlled. Qualified DDC users will request this report through The Chief of Research & Development, Department of the Army, ATTN: Director of Army Technical Information, Washington, D. C.

DISPOSITION INSTRUCTIONS

Destroy this report when it is no longer needed. Do not return it to the originator.

DISCLAIMER NOTICE

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPRODUCTION LIMITATIONS

No limitations.

30 November 1964

CIDS Report No. 3

A COLLECTION OF ALGORITHMS FOR SEARCHING CHEMICAL COMPOUND STRUCTURE ANALOGS

by

William J. Wilson
and
John B. Burger

General Electric Company
Computer Department
Huntsville, Alabama

Contract Number DA-01-021-AMC-242(Z)

Research Branch
Redstone Scientific Information Center
Directorate of Research and Development
U. S. Army Missile Command
Redstone Arsenal, Alabama

ABSTRACT

Some of the requirements of a computer system for searching chemical compound structure analogs are reviewed and algorithms are offered where appropriate. Among the factors discussed are file organization and data elements, the use of screens, the question of maximum query volume on a single pass of a master file, and the question of canonical forms. Of particular note are the experimental algorithms developed for the canonical ordering of finite undirected graphs permitting rapid determination of isomorphism between graphs to be accomplished.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Section I. INTRODUCTION

In any comprehensive computerized system for the performance of whole compound and fragment searches of a large file of chemical compound structure analogs, there are several major factors which must be carefully evaluated. Among these are:

1. The basic file organization and data elements.
2. The number, kinds, and fineness of screens employed.
3. The number of fragment queries which may be processed on one pass of the master file.
4. The desirability and feasibility of canonical forms for each compound structure analog.

A so-called "search algorithm" cannot effectively be developed without taking these factors into account. Because of this, a search algorithm will of necessity be based on the resolution of these questions and will itself embody many algorithms. Some of these would be concerned with the achievement of canonical forms, others with the logic of fragments searches, and still others with the application of screens. Each of the above factors will be discussed and algorithms offered where appropriate. In this manner, the reader may choose or discard algorithms as he sees fit or recombine them to suit the peculiar demands of a specific requirement.

Section II. FILE ORGANIZATION AND DATA ELEMENTS

Our system is postulated on the input availability of the following minimum elements of information for each compound:

1. BATCH Number[1]
2. Molecular Formula (optional, if not present it will be generated from the connectivity table. If present, it will be used as a redundancy check).
3. Arbitrarily numbered connectivity table containing the following elements of information:

Atom Number	Atom Qualification	Connectivity and Bond Qualification
01.	XX	NNBB, NNBB, NNBB, . .
02.	XX	NNBB, NNBB
03.	XX	NNBB, NNBB, NNBB. . .

(Where NN is the Number of the connected-to atom and BB is the value of the Bond)

From the above information, the following information will be generated:

1. Bond Summary*
2. Penny Connectivity Code for each atom[2]
3. The Level Coincidence Number if required (See Section "The Question of Canonical Form").

Each compound will be canonically ordered at input time and will be maintained in that form thereafter. The general logic for this operation is shown in Figure 1; the total Master Record Data elements are shown in Figure 2.

One of the basic decisions that must be made for any large-scale serial file is the basic order or orders of that file, and this decision

---

*Bond Summary is a tabulation of sums of each bond type; the bonds in the Bond Summary bear the same relation to Bond Summary as the atoms in a compound bear to the molecular formula.

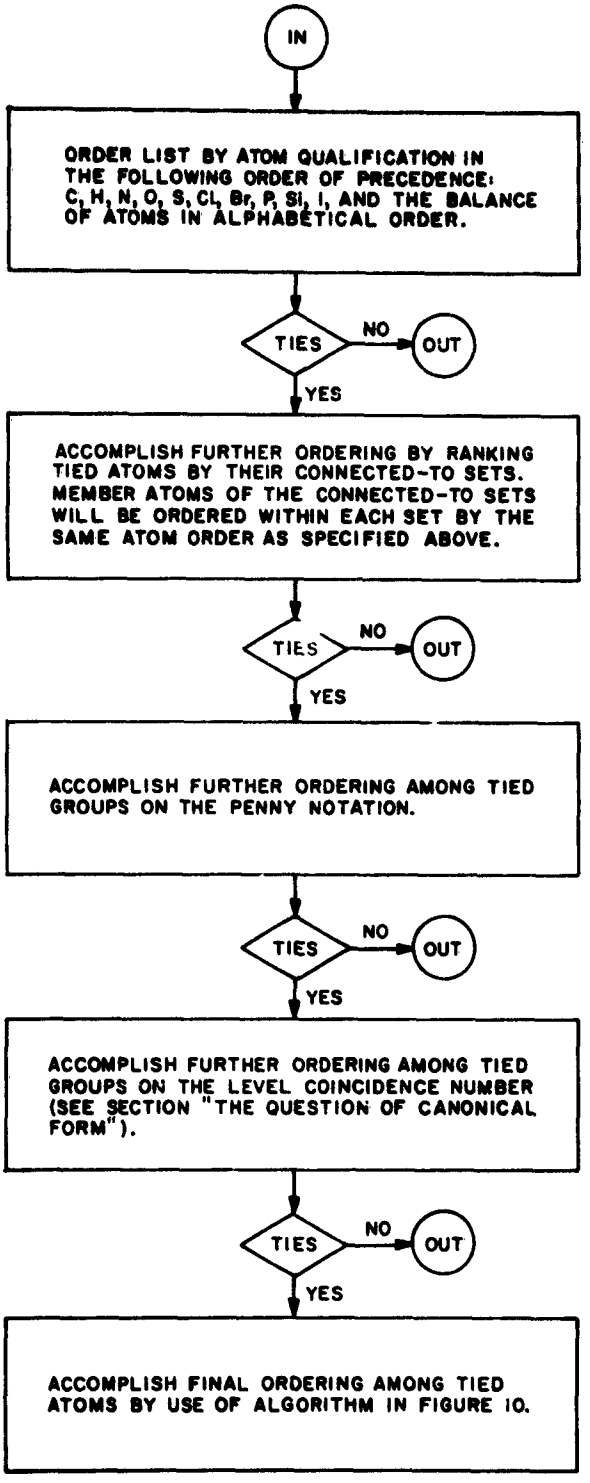

Figure 1. General Logic Diagram for Canonically Ordering Structural Analogs of Chemical Compounds

MASTER RECORD

Control Word Elements	Structural Data Elements				
BATCH Number Molecular Formula Bond Summary	Atom Number	Atom Identity	Connectivity and Bond Value	Penny Notation	Level Coincidence Numbers
	01	XX	NNB, NNB,...	/nn/nn/nn/	XXXX YYYYY
	02	XX	NNB, NNB	/nn/nn/	XXXX YYYYY
	03	XX	NNB, NNB,...	/nn/nn/nn/	XXXX YYYYY

Each Master Record will contain the above information for every Compound. The Control Word Elements are gross characterizations of the compound structure and are used as screens to preclude detailed searching where possible. This list is in canonical form. The Penny notation is capable of use as a fine screen if desired. This data is expanded, when required, for atom-by-atom and bond-by-bond searches.

Figure 2. Master Record Data Elements

is subject to many considerations.[3] Although many large printed compendia of chemical compound data are ordered by molecular formula or certain gross structural characteristics, magnetic tape files should not necessarily follow such an order. Wiswesser has shown that the distinctions afforded by molecular formula do not sufficiently distinguish the large majority of chemical compounds.[4] The need to search a file and to avoid as much as possible the detailed examination of each compound suggests a careful choice in the definitive index which must be appended to each compound. Such an index is found in the Wiswesser BATCH number. This does not, however, preclude the need for the molecular formula. In light of these and other considerations, we have constructed a control word for each compound (in the order presented) consisting of the BATCH Number, Molecular Formula, and Bond Summary.

The Master Compound File will be in major sequence by BATCH Number and within that by Molecular Formula and Bond Summary. The general logic for file maintenance and whole compound searching is outlined in Figure 3. For fragment searches, the general logic is outlined in Figure 4.

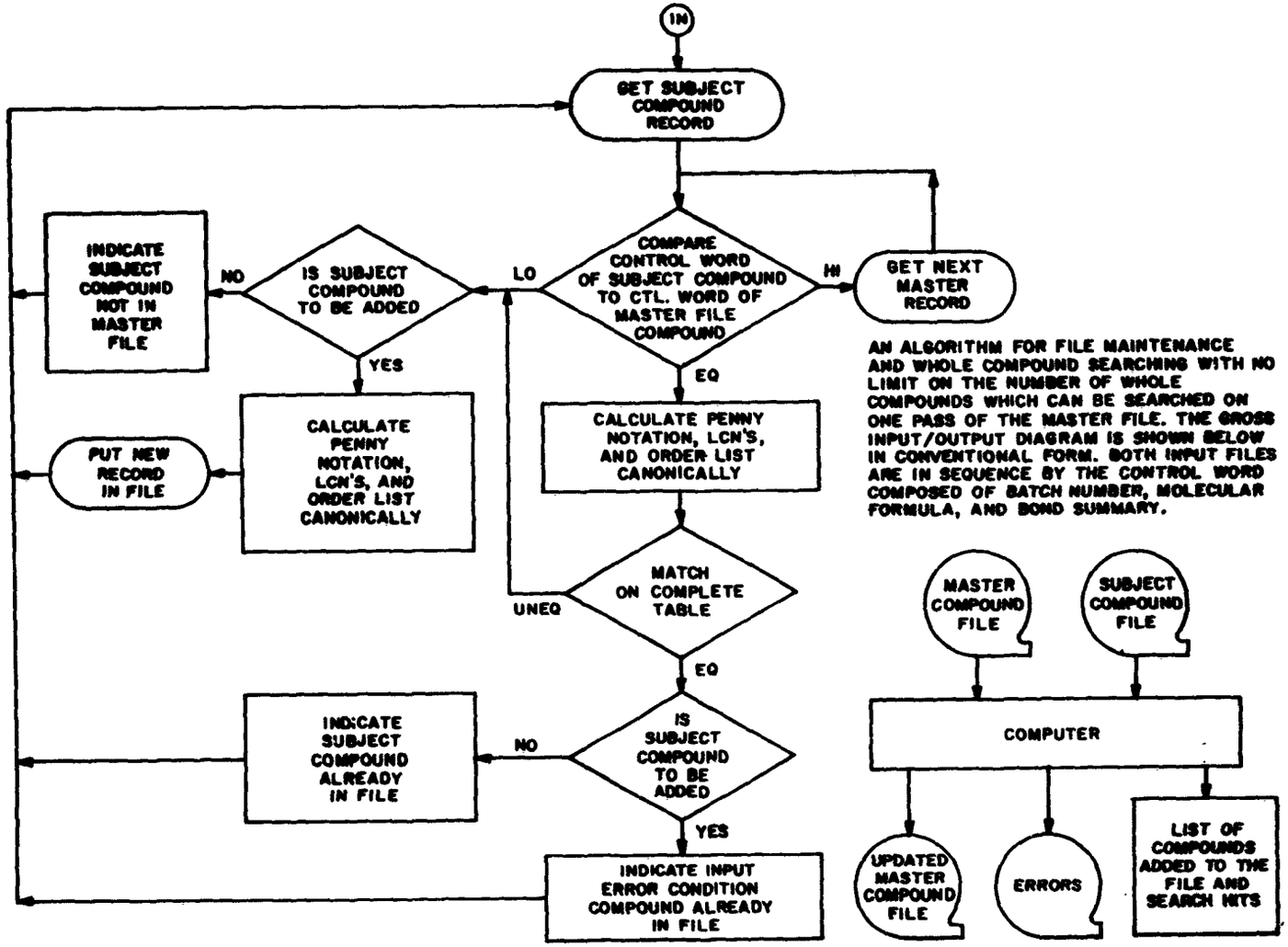

Figure 3. Algorithm for File Maintenance and Whole Compound Searching

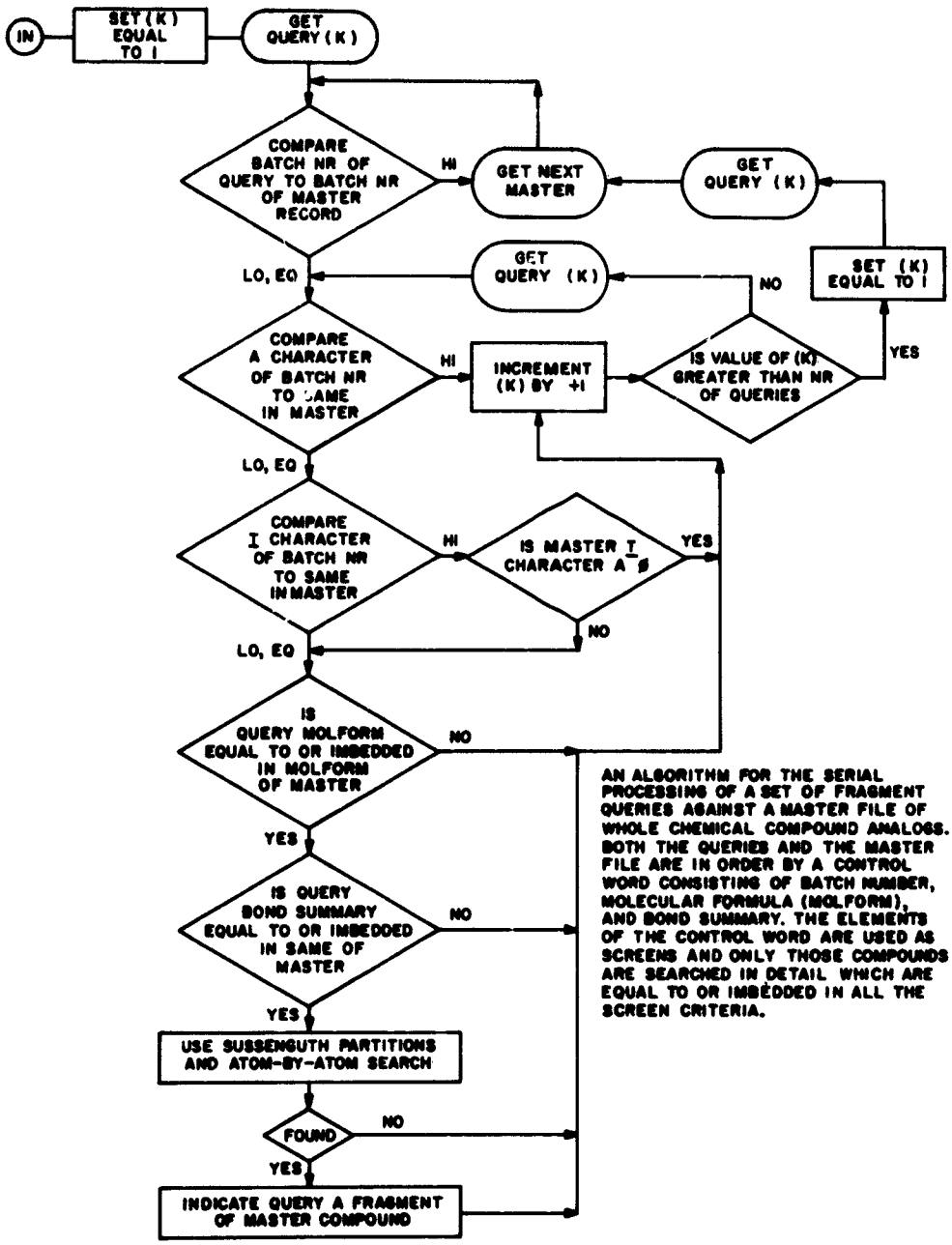

Figure 4. Algorithm for Serial Processing of Fragment Queries

Section III. SCREENS

The basic purpose of a screen is to preclude the detailed examination of each compound during a search. As such, the utilization of screens in a serial application is intimately bound up with the question of file sequence. In addition, the use of multiple screens offers a powerful tool in avoiding detailed searching when possible. A careful balance must be established empirically between the added time expended in screening compounds versus the time saved by avoiding detailed search. It is felt that considerable experimentation must be accomplished to determine the proper usage of screens.

Because of the proposed organization of our files using a single control word consisting of BATCH Number, Molecular Formula, and Bond Summary, a detailed examination of a given compound is required during whole compound searching only when it has the same control word as the query compound. When this is the case, a more detailed examination will be made utilizing the Penny connectivity codes. The complete logic for whole compound matching is shown in Figure 3.

A different approach is required for fragment searching. The fragment control word must be in all cases either identical to or imbedded in the master file control word before a detailed examination is required. The logic for fragment searching is shown in Figure 4.

Section IV. QUERY LIMITS

It is proposed that file maintenance and unlimited whole compound searching be accomplished simultaneously as shown in Figure 3. A separate operation is proposed for fragment searching, although there is no problem in incorporating whole compound searching with fragment searching if such is desired.

The question of how many fragment queries may be accomplished on a given pass of the master file is a serious one. In the terminology of information retrieval, the question (for a serial operation) is the choice of a linear file versus an inverted file. Although the use of inverted files places no limit on the number of queries that may be processed on one pass of the master file, the inversion of each chemical structure in terms of all its possible fragments reaches impractical proportions. On the other hand, the use of a linear file limits the number of queries as a function of the size of its fast-accessible store (memory, drum, etc.). If the number of queries exceeds the limit of the fast-accessible store, the queries must be placed on tape and the tape "seesawed" back and forth for each compound in the master file. At the present time, it is felt that a linear search should be accomplished with every effort expended to accommodate the largest number of queries. It was on this basis that the algorithms in Figures 3 and 4 were developed.

Section V. THE QUESTION OF CANONICAL FORMS

In many computerized, atom-by-atom, bond-by-bond searches (such as those employed by DuPont and the Chemical Abstracts Service), efforts have been directed toward the achievement of canonical forms for chemical compound structure analogs such that for each unique compound there is one and only one canonical form. This is usually accomplished by subjecting arbitrarily numbered structure analogs to a rigorous ordering procedure utilizing atom and bond qualifications and connectivity data as ordering criteria. The result is a unique list or matrix for each compound which is tantamount to a formal renumbering of the atoms of the subject compound.

The use of a canonical order greatly facilitates the process of whole compound matching. The same is not true for fragment searching, since the order accorded a fragment is not the same order accorded that fragment when it is a member of a large set. The evaluation of a total compound is accomplished in terms of all the member atoms, bonds, and connectivities of the subject compound. Thus "backtracking," which is the nemesis of fragment searching techniques, is required in many cases.

The question as to whether or not one may always achieve a canonical form for each graph analog of a compound appears still to be undecided. The presentation of counterexamples, such as the arbitrarily numbered planar graph in Figure 5 devised by Dr. Lehman of the Walter Reed Army Institute of Research, have necessitated revisions to canonical ordering techniques such as those used by DuPont and the Chemical Abstracts Service.

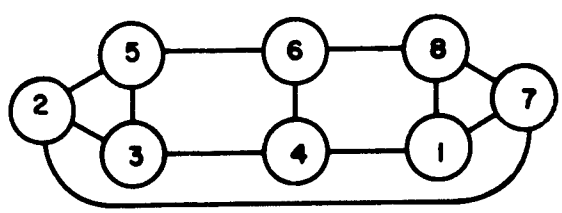

Figure 5. Lehman Counterexample

In light of this and other similar examples, it appears that a rigorous formal treatment is needed to decide whether the achievement of canonical forms of graph analogs of chemical compounds is always possible, or whether the question is truly undecidable.

Often, the test of a canonical ordering scheme is its ability to handle simple undirected graphs for which atom and bond differentiation is not available for ordering criteria, typified by the Lehman example. Such a test often unmasks latent fallacies in canonical ordering procedures which are dependent on the existence of atom and bond qualifications. The fact remains, however, that chemical compounds do possess such qualifications, and whether or not the inability of chemically oriented canonical ordering schemes to handle unqualified graphs is a real shortcoming remains to be seen.

Penny, in a recent paper[2], recognizes correctly that atom and bonding considerations alone are in some cases inadequate for distinguishing compounds. His method is concerned with enumerating the simple connectivity in the neighborhood of each atom. As he states, "it is a unique expression of the atomic network within the immediate neighborhood of the subject atom and is an attribute of the atom as much as its chemical indentity".

Returning to the Lehman graph (Figure 5), we can, by inspection, ascertain the topological equivalency of nodes 1, 3, 5, and 8, nodes 2 and 7, and nodes 4 and 6, which poses the following problems. Can this determination be accomplished algorithmically and, once done, can one rank the groups with respect to each other and, finally, can one rank the nodes within an equivalent set? If a general solution (algorithm) for achieving a canonical form can be derived for finite undirected graphs, there is no question of its applicability to the more highly differentiated graph analogs of chemical compound structures. Such a solution would provide the needed theoretical foundation for the development of canonical ordering routines for real compound structures.

The Lehman example rendered in the Penny notation yields the following table of connectivity codes:

1. /22/22/22/
2. /22/22/22/
3. /22/22/22/
4. /22/22/22/
5. /22/22/22/
6. /22/22/22/
7. /22/22/22/
8. /22/22/22/

The undifferentiated aspect of this table is more a peculiarity of the Lehman graph itself than a shortcoming of the notation. For this and all similar examples a more exhaustive approach must be undertaken.

Using the Penny concept (less the notation), each node of the Lehman graph is defined to a depth of three levels using the arbitrary node numbers shown in Figure 5. The resultant lists are shown diagrammatically in Figure 6. In dealing with connectivity tables of arbitrarily numbered graphs, it must be kept in mind that no significance attaches to the particular number assigned any node. However, an evaluation of each node in terms of all other nodes at a given level can be made to yield information, and this information can ultimately be used to canonically number graphs.

As can be seen in Figure 6, each node of the Lehman graph is depicted through three levels of connectivity. Each node list is analyzed according to the number of times a given node number repeats at each level, and this information is summarized in the "level coincidence number" or LCN. For node 1, for example, at the second level there is an LCN of 4100 which is interpreted as follows: there are four unique node numbers, one duplicate, and no triplicates or quadruplicates appearing at this level. The topological equivalency of nodes 1, 3, 5, and 8 in the Lehman example is reflected in the equivalency of their respective level coincidence numbers at the second and third level, which are 4100 and 4201 in all four cases. Equivalent nodes 4 and 6 have level coincidence numbers of 2200 and 4120 at the second and third level, and equivalent nodes 2 and 6 have level coincidence numbers of 6000 and 6000 at the second and third level.

Further experiments have demonstrated the utility of the level coincidence number as a measure of the degree of connectivity or "imbeddedness" of the nodes of a graph. As a result, we have identified three classes of graphs:

Class 1 - Differentiated, i.e., a graph in which each node enjoys a unique degree of imbeddedness.
Class 2 - Mixed, i.e., a graph in which some nodes enjoy the same degree of imbeddedness.
Class 3 - Undifferentiated, i.e., a graph in which every node enjoys identical imbeddedness.

Shown below in Figure 7 are three graphs which demonstrate respectively the Class 1, Class 2, and Class 3 distinctions. It should be noted that graphs B and C obtain different classes even though both are eight-node graphs regular of degree three.

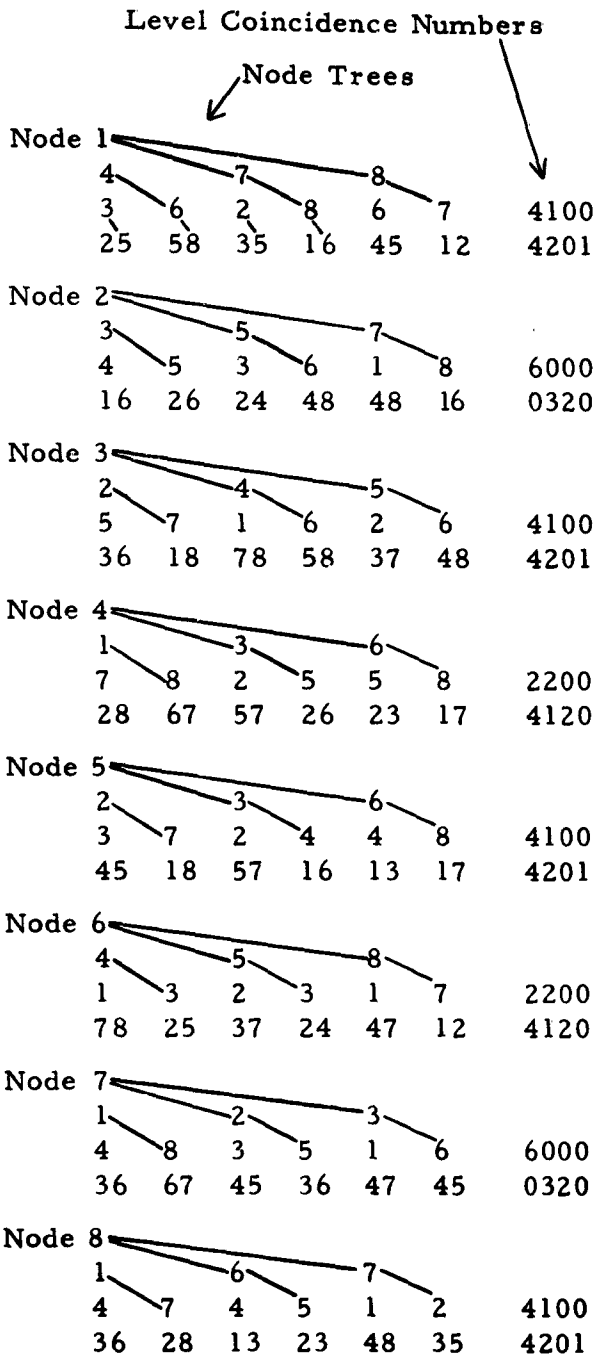

Figure 6. Level Coincidence Number Relations

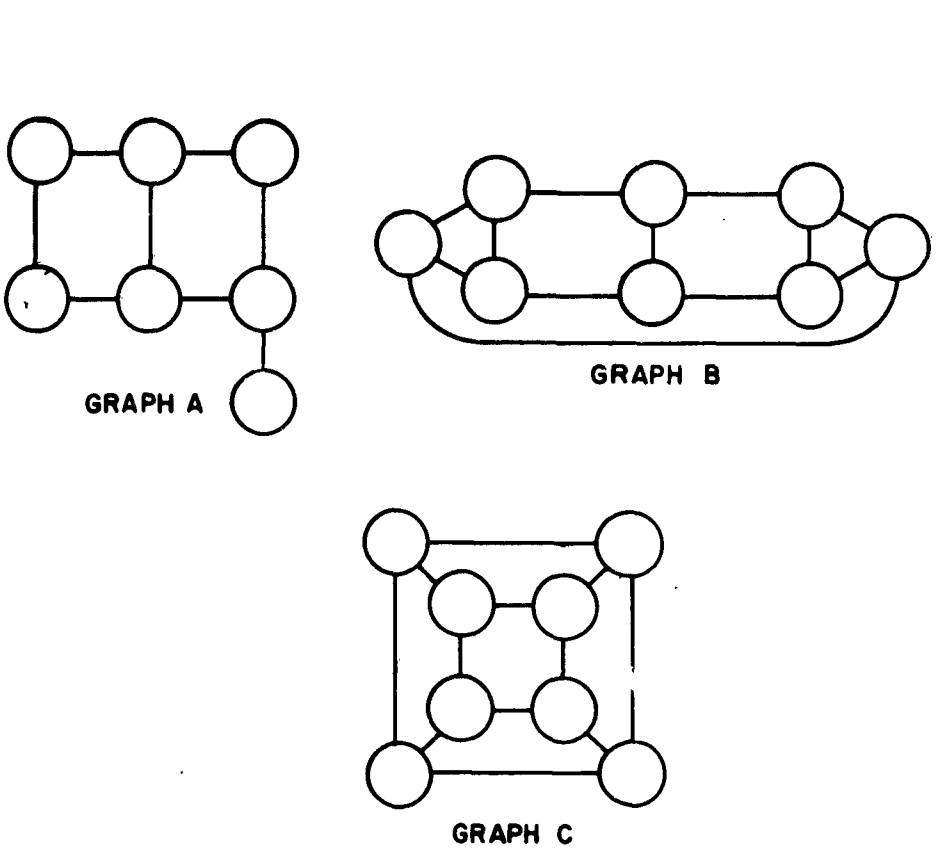

Figure 7. Examples of Basic Graph Types

The development of the LCN does not, in itself, solve the problem of canonically ordering the nodes of a graph since there may be duplicate LCN's (Class 2) or the graph may consist entirely of nodes with equivalent LCN's (Class 3). For these cases, the experimental algorithms in Figures 8 and 9 have been devised. The algorithm in Figure 9 is concerned solely with canonically ordering Class 3 graphs and resolving ties in Class 2 graphs. The algorithm in Figure 8, using the alg rithm in Figure 9 as a subroutine, is addressed to the problem of canonically ordering any finite undirected graph without loops and multiplicity edges.

Shown below in Figure 10 are the planar representations of two Class 3 graphs representing the nodes of a cube and a dodecahedron, respectively. Both of these have been canonically numbered using the algorithm in Figure 8. These and other Class 3 examples have been tested by furnishing the unnumbered graphs and the algorithms to clerks for numbering and verifying their equivalency.

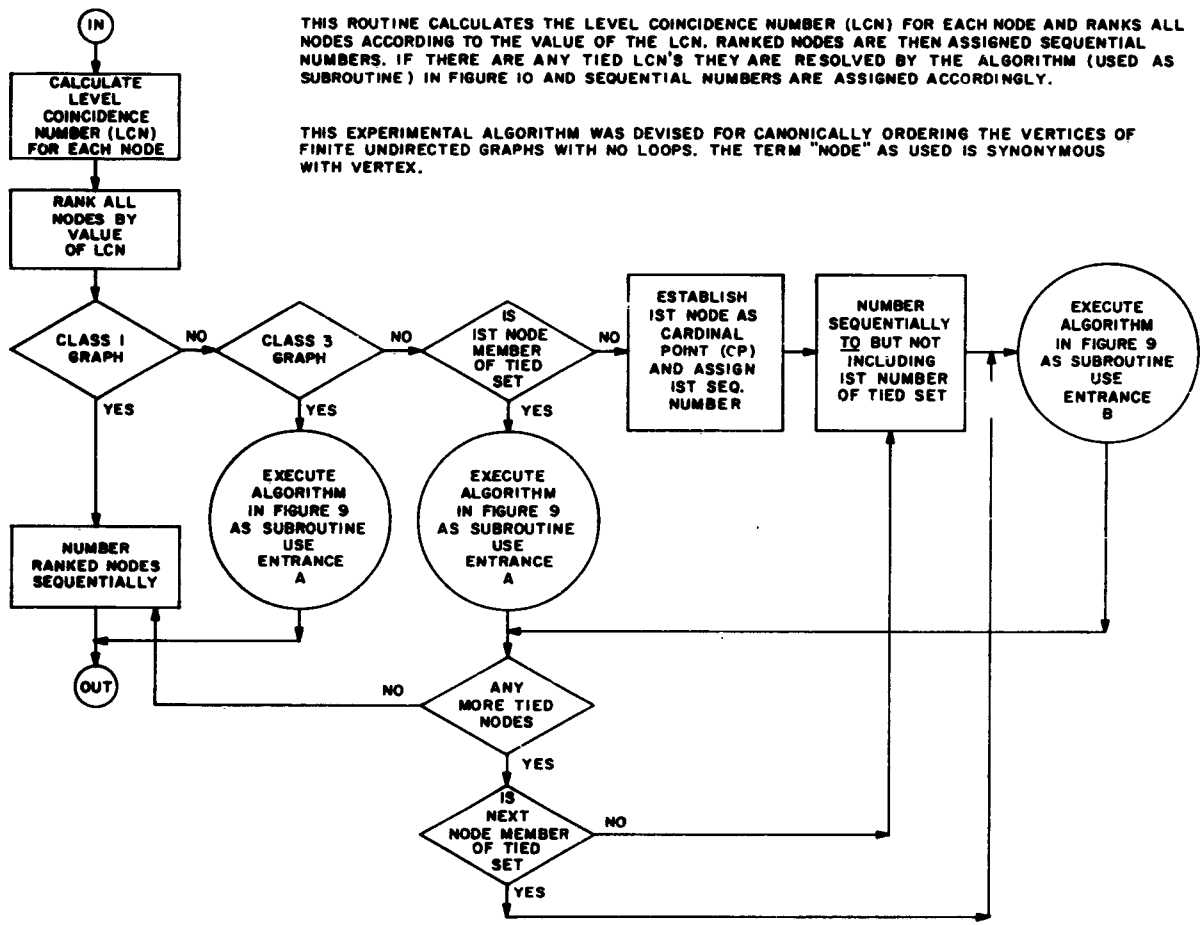

Figure 8. Experimental Algorithm for Canonically Ordering Finite Undirected Graphs With No Loops

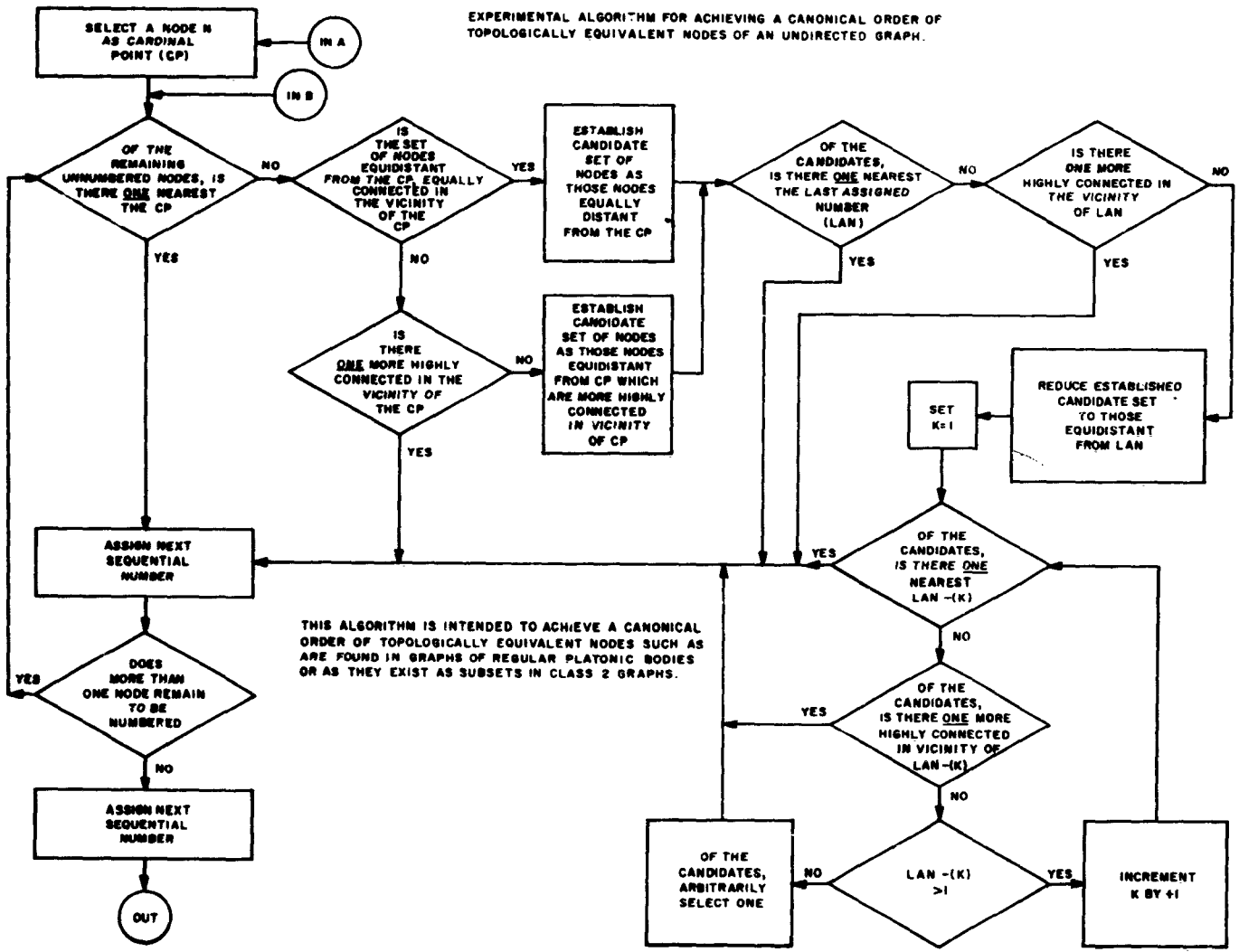

Figure 9. Experimental Algorithm for Canonically Ordering Topologically Equivalent Nodes of Finite Undirected Graphs With No Loops

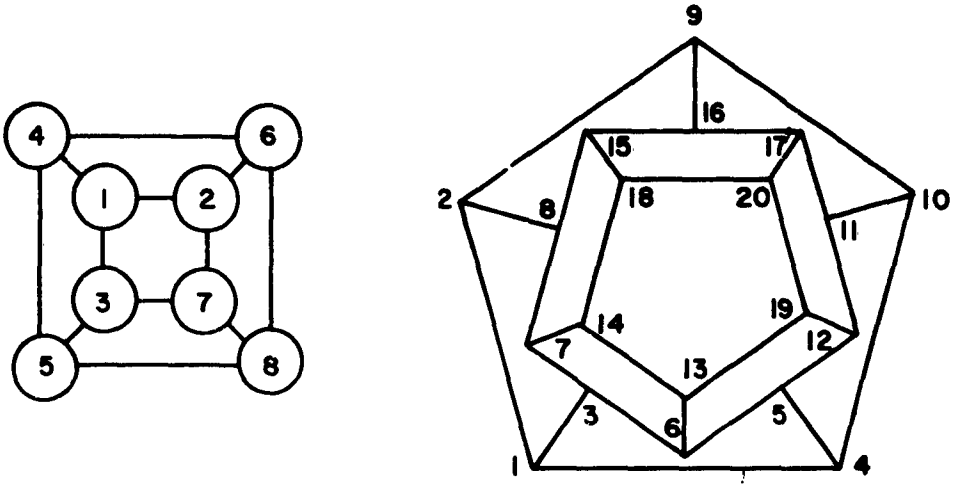

Figure 10. Two Class Three Graphs

Section VI. OTHER CONSIDERATIONS

Results obtained thus far on the problem of graph isomorphism suggest further investigation of the following considerations:

1. Given a graph of n-nodes, what is the maximum level to which one must descend in the development of LCN's to determine the existence of a Class 3 graph?

2. In the iteration in Algorithm 7 to decide which node is nearest the node LAN - (k), is there a limit on (k)? Experience, thus far, indicates that a selection is made at either LAN - (1) or LAN - (2) or [illegible] lection is made, LAN -(k) goes to 1, and the selection is arbitrary.

3. In the development of the LCN's, what is the highest n-tuplicate one may expect at a given level for a graph of n-nodes?

4. In some respects, the procedure for numbering Class 3 graphs is analogous to removing nodes as they are numbered. Can an algorithm be devised operating on this principle of removing nodes (and edges) and numbering the remaining graph?

5. The relative "imbeddedness" of nodes may be ascertained by tabulating the number of numbers appearing at a specified level. Is this analogous to finding the center (s) of a graph? This information is carried in the LCN and is derivable by the following computation: given an LCN of 4201 it yields an imbeddedness value calculated as follows (4x1) + (2x2) + (0) + (4x1).

6. Can a simple method be devised through analysis of raw level numbers (or their LCN's) to determine if a graph is planar or nonplanar?

LITERATURE CITED

1. William J. Wiswesser, CWIK LIST News, Chemical World Index Key, 3103 River Road, Reading, Pa., December 1963.

2. Robert H. Penny, "A Connectivity Code for use in Describing Chemical Structures," General Electric Company, Computer Department, June 9, 1964.

3. J. B. Burger and William J. Wilson, "A Review of Selected Methods of Machine Manipulation of Chemical Structures," General Electric Company, Computer Department, September 1964, AD451700.

4. William J. Wiswesser, "A Line-Formula Chemical Notation," Thomas Y. Crowell, 1954.

5. Sylvan Eisman, Private Communication, June 13, 1964.

6. E. H. Sussenguth, "A Graph-Theoretic Algorithm for Matching Chemical Structures," The Computation Laboratory of Harvard University.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY *(Corporate author)*
Redstone Scientific Information Center
Directorate of Research and Development
U. S. Army Missile Command
Redstone Arsenal, Alabama

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b GROUP
N/A

3 REPORT TITLE

A COLLECTION OF ALGORITHMS FOR SEARCHING CHEMICAL COMPOUND STRUCTURE ANALOGS

4 DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Final Progress Report

5 AUTHOR(S) *(Last name, first name, initial)*

Wilson, William J. and Burger, John B.

6. REPORT DATE	7a TOTAL NO OF PAGES	7b NO OF REFS
November 1964	19	6

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
DA-01-021-AMC-242(Z)	
b. PROJECT NO. 4	CIDS-3
c	9b OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d	AD

10. AVAILABILITY/LIMITATION NOTICES

Qualified requesters may obtain copies of this report from DDC.

11. SUPPLEMENTARY NOTES	12 SPONSORING MILITARY ACTIVITY
None	Same as 1

13. ABSTRACT

Some of the requirements of a computer system for searching chemical compound structure analogs are reviewed and algorithms are offered where appropriate. Among the factors discussed are file organization and data elements, the use of screens, the question of maximum query volume on a single pass of a master file, and the question of canonical form. Of particular note are the experimental algorithms developed for the canonical ordering of finite undirected graphs to permit the determination of isomorphism between graphs.

DD FORM 1 JAN 64 1473

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Chemistry - Information Retrieval Information Retrieval - Computer Chemical Structure Analogs - Computer Search Algorithms - Chemical data Graph Isomorphism Determination Canonical Forms - Achievement of File Organization - Serial processing Data Screens Query Volumes - The question of maximum						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13 ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, *or (U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.